Journal of Soft Computing in Civil Engineering 8-3 (2024) 28-48

Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Predictive Performance Analysis of NLP-Based Knowledge Summarization in High Entropy Alloys (HEAs) Literature

Akshansh Mishra [1*]; Vijaykumar S Jatti [2]; Vaishnavi More [3]; Eyob Messele Sefene [4]; Anish Dasgupta [5]; Devarrishi Dixit [6]

1. School of Industrial and Information Engineering, Politecnico Di Milano, Milan, Italy
2. Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Pune 412115, India
3. Software Developer, Turabit LLC, Ahmedabad, Gujarat 380052, India
4. Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taiwan
5. Software Engineer, ALTEN India, Bangalore, India
6. Department of Materials Science Engineering, Christian Albrechts University zu Kiel 24143, Germany

Corresponding author: *eyob.messele@bdu.edu.et*

https://doi.org/10.22115/SCCE.2024.383343.1599

ARTICLE INFO

Article history:
Received: 26 January 2023
Revised: 23 June 2023
Accepted: 10 January 2024

Keywords:
Natural language processing;
Artificial intelligence;
High entropy alloys;
Knowledge summarization.

ABSTRACT

The ability to understand spoken language is closely tied to natural language processing (NLP) in AI systems. NLP aims to teach machines how words are connected, used in context, and understood to mimic human cognition. By leveraging NLP, AI systems can quickly comprehend spoken and written language, surpassing human processing capabilities. In this research paper, we focus on five NLP algorithms—Lex Rank, Sumy, Luhn, Latent Semantic Analysis (LSA), Kull-back-Liebler (KL) algorithm, and Text Rank—for knowledge summarization in the field of High Entropy Alloys (HEAs). We evaluate the performance of these algorithms using the BLEU score and ROUGE score as metrics. Our findings indicate that the Luhn algorithm achieves the highest accuracy in knowledge summarization, outperforming the other algorithms used in this study.

How to cite this article: Mishra, Jatti VS, More V, Sefene EM, Dasgupta A., Dixit, D. Predictive Performance Analysis of NLP-Based Knowledge Summarization in High Entropy Alloys (HEAs) Literature. J Soft Comput Civ Eng 2024;8(3):28–48. https://doi.org/10.22115/scce.2024.383343.1599

2588-2872/ © 2024 The Authors. Published by Pouyan Press.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Natural language processing is the ability of a computer program to interpret natural language, or a system of communication, in both its spoken and written forms (NLP). It falls under artificial intelligence (AI), and Linguistics originates in NLP, which has been around for over 50 years. It is helpful in various fields, including search engines, corporate intelligence, and medical research. [1–3]. NLP has made it possible for computers to understand natural language in the same way that humans do. Whether the speech is written or spoken, natural language processing uses artificial intelligence to gather direct experience, analyze it, and organize it logically so that a computer can understand it. Similar to how individuals have multiple sensors like ears to listen and eyes to see, computers have reading programs and speakers to gather sounds. Systems have a script to process their various inputs, just as individuals have a mind to do so. The input is eventually translated into computer-readable code during processing [4,5]. Businesses need a way to effectively process the vast amounts of unorganized, text-heavy data they use. Until recently, companies could not efficiently analyze the natural human language that made up a large portion of the information generated online and kept in databases. Natural language processing comes in handy in this situation.

Consider the following two sentences: “Cloud computing coverage should be included in every service-level agreement” and “A solid SLA guarantees an easier night’s sleep-even in the cloud” to see the benefit of NLP. Suppose a user uses NLP to search. In that case, the software will understand that cloud computing is a concept, that cloud is an acronym for cloud computing, and that SLA is an abbreviation for service-level agreement. For various purposes, including data gathering, hypothesis building, and pattern detection within and across domains, NLP can be employed for materials, as stated by Olivetti et al. [6] and illustrated with instances. Tshitoyan et al. [7] employed an unsupervised method for materials. The earlier articles contained significant implicit knowledge about upcoming findings. This further demonstrates that earlier articles include substantial implicit knowledge about forthcoming findings. Their results, which hint at a complete technique for mining scientific literature, emphasized the possibility of jointly retrieving information and connections from a large amount of scientific literature. Mohammadi et al. [8] used 1120 articles published between 1974 to 2007 and evaluated the transdisciplinary tendencies of Iranian research in NST. Using text mining tools, it was discovered that the fundamental ideas of the Iranian literature included in NST were contained in 96 phrases. By using multidimensional scaling and basing it on the co-occurrence of the essential phrases in the academic articles, the scientific backbone of the Iranian NST was then reconstructed. The findings demonstrated that the multidisciplinary structure of the NST domain in Iranian papers includes pure physics, analytical chemistry, material science and engineering, biochemistry, chemical physics, physical science, and new developing themes. The analysis indicates that whereas most RMC literature has been addressed to concrete technology and material science, very few studies have concentrated on RMC dispatching [9].

Learning from past mistakes is necessary to find the root reasons and pertinent elements. A valuable learning resource is the enormous corpus of textual data on event narratives accumulated over time. Due to the massive amount and chaotic nature of the text data, it is

difficult to deduce recurrent recurrence trends from it. Using the pipeline industry as an example, Liu et al. [10] used NLP and textual mining algorithms to exploit the resource and comprehend the occurrences' underlying causes and contributing aspects. The Pipeline and Hazardous Materials Safety Administration's (PHMSA) incident database's 3587 event narrative entries were used in the "comment" section. It has been shown from prior research articles that only a few studies have used NLP in the field of material science to achieve knowledge summarization goals. Therefore, this is the first research study to use NLP for summarizing knowledge on High Entropy alloys [11,12,21–30,13–20]. The future scope of this work is to implement this approach in other domains of material science for knowledge extraction purposes [31–36].

The main aim of this research is to assess the performance of five distinct natural language processing (NLP) algorithms—Lex Rank, Text Rank, Luhn, Latent Semantic Analysis (LSA), and Kull-back-Liebler (KL) algorithm—in the context of knowledge summarization for High Entropy Alloys (HEAs). The objective is to evaluate the accuracy and efficacy of these algorithms by utilizing evaluation metrics such as the BLEU score and ROUGE score. The specific goal is to identify the algorithm that demonstrates the highest level of performance in summarizing HEAs literature. By accomplishing this objective, the research endeavors to make valuable contributions to the field of NLP and advance the automation of summarizing technical literature pertaining to High Entropy Alloys.

2. High entropy alloys and their applications

In a multi-component system, high entropy alloys (HEAs) are next-generation alloys with multiple primary elements. MPEAs (multi-principal element alloys) and complex, concentrated alloys are other names for HEAs (CCAs). Professor Ye Junwei first put forth the idea of high entropy alloys (HEAs) and multi-principal element alloys (MPEAs) in 2004 [37,38]. Since each HEA can be adjusted by small elemental additions, like with present element-based alloys, each HEA is a new alloy base. Numerous novel alloy bases are provided by HEAs [39]. Equation 1 provides the number of HEA systems (unique combinations of elements without disclosing composition).

$$C\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{1}$$

where n is the number of components in the array used to choose the r primary alloy elements. Researchers participating in alloy creation historically concentrated on the phase diagram's corners to create a conventional alloy, which only takes up a small part of the design space (see Figure 1). However, the recent advancement of HEAs has caused the center area to receive more attention [40].

Every one of the at least five metallic materials used to create high entropy alloys (HEAs) has an atomic concentration that ranges from 5 to 35% in the molar form. High entropy alloys have great work hardenability, high-temperature oxidation resistance, and good ductility. Additionally, HEAs have outstanding attractive magnetic characteristics, high wear resistance, and superior erosion resistance. The three main categories of HEA fabrication techniques are

liquid mixing, solid mixing, and gaseous mixing. Arc melting, electric resistance melting, inductive melting, Bridgman solidification, and laser additive manufacturing are all types of liquid mixing [41–46].

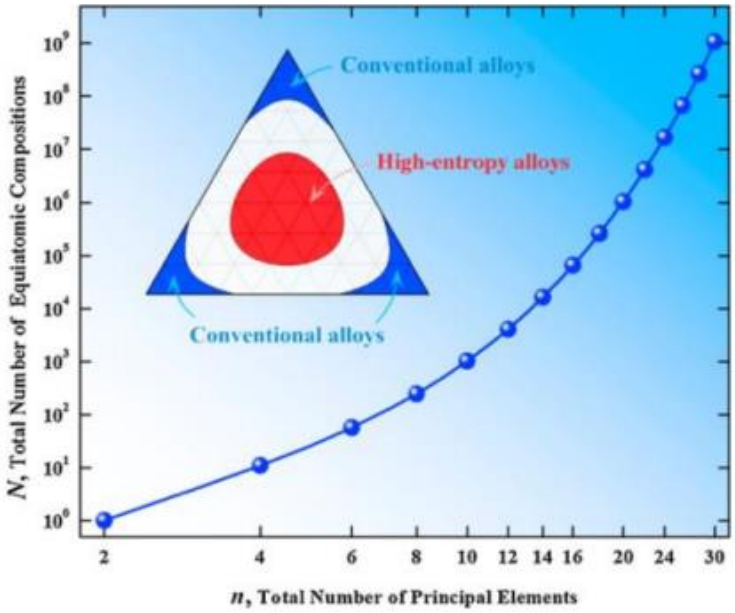

Fig. 1. The relationship between the total number of primary elements and the total number of equiatomic compositions. The ternary figure in the inset shows how the design of high-entropy alloys differs from that of normal alloys [40].

Figure 2 indicates four categories of entropic alloys: high entropy alloys, medium entropy alloys, low entropy alloys, and pure metal. When it comes to low-temperature applications, high-entropy alloys perform better than conventional materials.

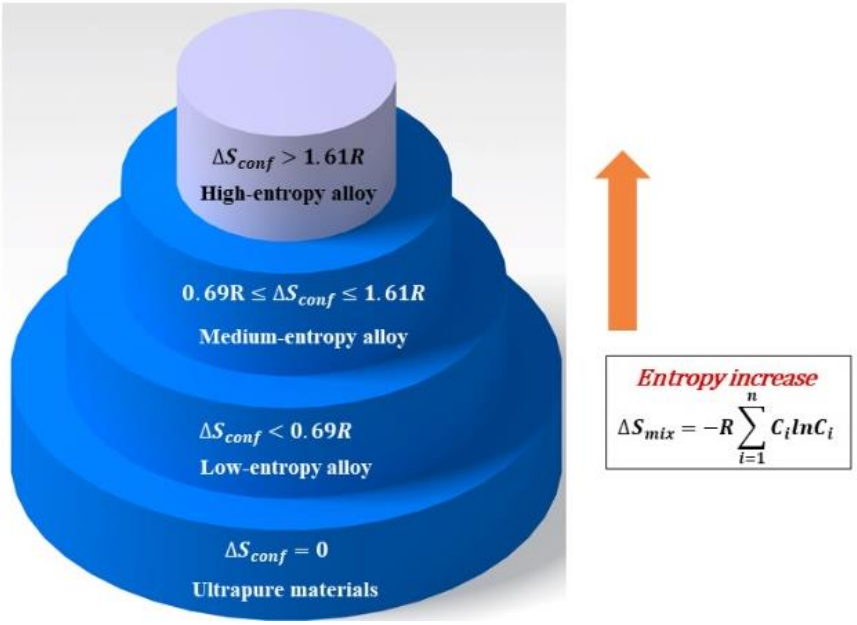

Fig. 2. Types of High Entropy Alloys [47].

Cryogenic processes use alloys with high entropy [47]. They can be employed as low-temperature materials in the civil engineering, superconducting, and aircraft sectors. Geanta et al. [48] tested and characterized High-Entropy Alloys from the AlCrFeCoNi System for various military applications. According to microstructure research, the microstructure of alloys with high entropy appeared frozen when they were molten, as illustrated in Figure 3. As illustrated in Figure 4, impact testing results revealed that high entropy alloy is the best choice for ballistic packages robust at high-velocity penetration impacts.

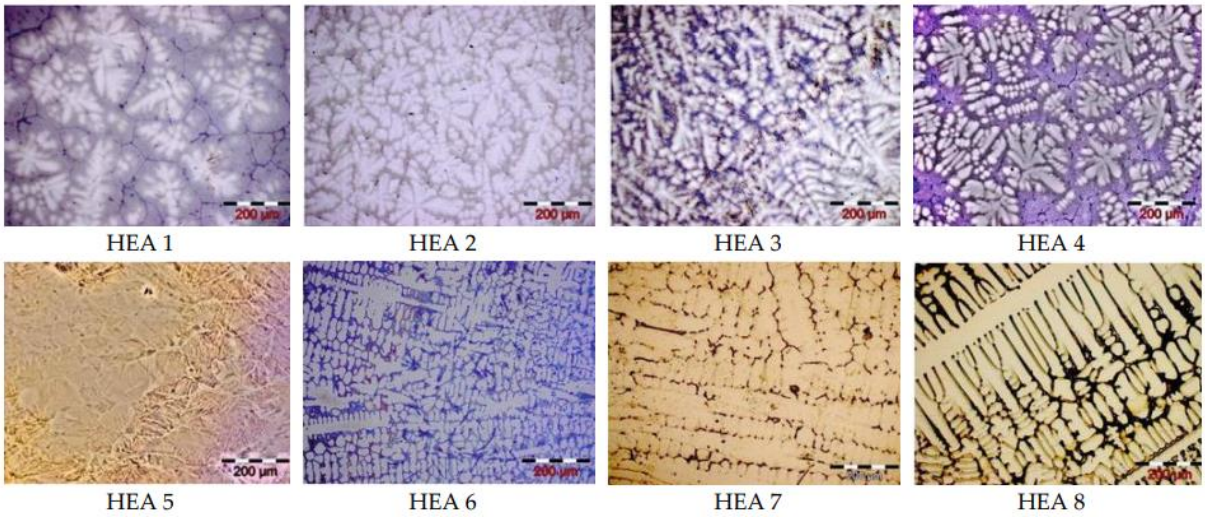

Fig. 3. Microstructure analysis of experimental as-cast AlCrxFeCoNi alloys [48].

Fig. 4. Dynamic testing on a HEA-steel ballistic package (a) after the first fire and (b) just after the second fire [48].

HEAs research involves intricate materials science concepts, experimental data, alloy compositions, and properties. Summarizing this wealth of information helps in distilling the crucial aspects and core principles underlying HEAs. By providing concise summaries, knowledge summarization enables researchers and stakeholders to comprehensively understand HEAs, their characteristics, and potential applications more efficiently.

3. Understanding natural language algorithms

Computer science's field of natural language processing (NLP) investigates how computers and human languages interact. It is the engine that powers search engines like Google. Manual linguistic analysis is possible and has been practiced for centuries. But technological advancements continue, particularly in natural language processing (NLP). Natural Language Processing (NLP) has been aggressively pursued by the machine and deep learning fields using various methods. Even though some of the methods we use today have only been around briefly, they are already altering how we communicate with technology. Research in natural language processing (NLP) shows us how to create systems that can comprehend human language in practical ways. Among many others, these include chatbots, machine translation software, and speech recognition systems. The same problem, however, can be solved using various algorithms. This article compares four widely used techniques for teaching ma-chine learning models to analyze human language data.

NLP (natural language processing) algorithms come in several forms. They can be divided into groups based on their tasks, such as entity recognition, relation extraction, entity parsing, and part of speech tagging. The tags created by Part of Speech Tagging algorithms describe how specific phrase components work. For instance, the usual tags for words in English phrases are verbs, nouns, and prepositions. The parsing process breaks down a sentence into its grammatical constituents (phrase structure), typically by splitting it into phrases and pieces of speech that trees can represent. Named entities, such as persons, companies, or locations, are found within a text using named entity recognition. Semantic analysis's subsidiary task is relationship extraction. According to relation extraction, there must be a connection between two items if they may be connected by an auxiliary verb (such as "is," "was," or "was not," etc.) or another marker word (such as "to" or "from").

Several NLP algorithms are created with various goals in mind, from language production to sentiment analysis. One technique for helping computers comprehend your intentions when you speak or write is sentiment analysis. Companies employ sentiment analysis as a tool to ascertain whether their customers are satisfied with the goods or services they provide. However, it also has the potential to better understand how individuals feel about issues in politics, the medical field, or any other field where individuals have different convictions. Based on their language and social media postings and comments, sentiment analysis may tell whether someone is feeling positively or negatively about anything. Sentiment analysis is important because it enables businesses to handle dissatisfied clients more effectively. Named entity recognition is sometimes thought of as text classification, where a group of documents must be categorized according to person or organization names. Algorithms for machine learning typically handle this task. Many classifiers are available, but the k-nearest neighbor technique is the most straightforward (kNN). This straightforward technique is predicated on the idea that numerous words adjacent to a specific thing described in a document are also parts of that entity. Text processing tasks like text summarization have received a lot of attention recently. The fundamental goal of text summarizing is to provide an edited version of the initial text that only conveys the key ideas. The two types of text summarization are extractive and abstractive. Term frequency-inverse

document frequency (TF-IDF) score, frequency words, and other statistical measures linked to centrality (i.e., how significant they are) are used to choose sentences from the source documents in the extract text summarization method, also known as key phrase extraction. Abstractive text summarization produces new phrases with details not found in the source text. A subset of natural language pro-cessing is called aspect mining. Aspect mining identifies the various traits, components, or facets of text. Aspect mining divides texts into various categories in order to pinpoint the attitudes—often referred to as sentiments—described in each category. The topic is classified rather than the feeling when aspects and subjects are compared. Depending on the technique employed, aspects might include things, acts, feelings or emotions, qualities, occurrences, and more. Companies have used aspect mining methods to identify customer responses. Aspects can be important components in the research and development of artificial intelligence for new technologies like chatbots, which use AI programs to respond with appropriate answers based on information gleaned from text conversations that they analyze for patterns and interconnection between questions and answers. To obtain explicit or implicit attitudes regarding features of text, aspect mining is sometimes integrated with sentiment analysis tools, a different sort of natural language processing. Because aspects and views are so closely related, they are frequently employed as synonyms in literature. Companies might benefit from aspect mining since it enables them to identify the type of client replies.

4. Used natural language processing in the present work

4.1. Text rank

The TextRank algorithm is a powerful natural language processing (NLP) technique employed for automated text summarization and keyword extraction. Inspired by Google's PageRank algorithm, TextRank follows a similar principle: rank sentences or words based on their significance within a text. The mechanism of the TextRank algorithm involves several key steps. First, the input text is segmented into individual sentences to allow for independent analysis. Next, the sentences undergo preprocessing steps such as tokenization, stop word removal, and part-of-speech tagging to ensure meaningful analysis. A graph is then constructed, with each sentence represented as a node and edges connecting nodes based on semantic similarity or co-occurrence. The algorithm proceeds to iteratively rank the sentences by assigning initial equal scores and updating them based on the scores of connected sentences. The update process considers factors like similarity and importance. The iterative ranking continues until convergence is reached. Finally, a threshold is applied to select the most important sentences for the summary. With the highest scores, these selected sentences are deemed the most relevant and informative. The underlying principle of TextRank is that important sentences tend to be connected to other important sentences within the text. By leveraging the graph-based representation and iterative ranking, TextRank effectively identifies and highlights the most significant sentences for summarization. With its ability to automatically summarize text, TextRank offers a valuable tool for extracting key information from large volumes of text data, aiding in efficient information retrieval and decision-making processes.

4.2. Lex rank

LexRank is a noteworthy algorithm utilized in natural language processing (NLP) for text summarization. It operates on the principle that sentences carrying more similarity to other sentences within a document are likely to be more significant. The mechanism of LexRank involves several steps. Initially, the input document is segmented into individual sentences. These sentences then undergo preprocessing, including tokenization, stop word removal, and part-of-speech tagging. Next, the algorithm calculates the similarity between sentence pairs using metrics like cosine similarity or Jaccard similarity. This information is used to construct a graph representation, where sentences are nodes and edges indicate similarity scores. The algorithm applies eigenvector centrality to rank the sentences within the graph, considering both a sentence's own score and the scores of its connected sentences. The iterative process continues until convergence is reached. Finally, a threshold is applied to select the most important sentences for the summary based on their scores. LexRank's approach effectively captures the global structure of the document by evaluating sentence relationships. This graph-based technique, coupled with eigenvector centrality, enables the identification of key sentences that convey essential information. LexRank provides a robust and scalable solution for automatic text summarization, facilitating extraction of crucial content from extensive textual data.

4.3. Luhn algorithm

The Luhn algorithm is a widely used technique in natural language processing (NLP) for text summarization and keyword extraction. It operates based on the idea that important words or phrases in a document often occur frequently. The mechanism of the Luhn algorithm can be summarized as follows. First, the input document undergoes preprocessing steps, including tokenization and stop word removal, to prepare it for analysis. Then, the algorithm assigns a score to each word or phrase based on its frequency of occurrence. However, common words are given lower scores to avoid bias toward them. The algorithm also takes into account the proximity of important words to each other, giving higher scores to words that are close in the text. After scoring all the words or phrases, a threshold is applied to select the most significant ones for the summary. These selected words or phrases contribute to the overall understanding of the document. The Luhn algorithm's simplicity and effectiveness make it popular for keyword extraction and text summarization tasks. Identifying important words or phrases based on frequency and proximity helps distill the essential content from a document and aids in generating concise summaries or extracting key information.

4.4. Latent semantic analysis

Latent Semantic Analysis (LSA) is an NLP technique that uncovers hidden semantic relationships and meanings within a text collection. It operates on the principle that words appearing in similar contexts will likely have similar meanings. LSA involves creating a term-document matrix to capture term occurrences across documents. The matrix is transformed through Singular Value Decomposition (SVD), revealing latent semantic factors that represent the underlying connections between terms and documents. By comparing these factors, LSA determines the semantic similarity between terms and documents. LSA has applications in

information retrieval, document classification, and text summarization, as it provides a deeper understanding of the relationships between words and documents, facilitating more accurate analysis and interpretation of text data.

4.5 Kull-back-Liebler (KL) algorithm

The Kullback-Leibler (KL) algorithm, also called Kullback-Leibler divergence, is a statistical technique employed in NLP for tasks like information retrieval and text classification. It calculates the difference between two probability distributions, usually representing a reference and a target distribution. By comparing the distributions, the KL algorithm measures the information loss when approximating the target distribution with the reference distribution. It provides insights into how well the reference distribution represents the target distribution. The KL algorithm finds applications in text summarization, information extraction, and language modeling, enabling effective textual data analysis and comparison by capturing distributional differences or similarities.

5. Metrics features for evaluating the performance of knowledge summarization in the present study

In the present work, we have implemented metrics features such as BLEU and ROUGE for measuring the performance of NLP algorithms for knowledge summarization purposes. Both the ROGUE and BLEU sets of metrics can be used to create text summaries. BLEU was first required for machine translation, but it is as suitable for the task of text summarization.

A candidate translation of a text is scored using the Bilingual Evaluation Understudy (BLEU), which compares it to one or even more reference translations. Despite being designed for translation, it has the potential to assess text produced for a variety of NLP activities. The initial stage is the computation of Precision scores for the 1- to 4-gram weight range stated in equations 2.

$$Precision\ of\ N-Grams\ (p_n) = \frac{Number\ of\ correct\ predicted\ n-grams}{Number\ of\ total\ predicted\ n-grams} \tag{2}$$

Next, we use Equation 3 to aggregate these Precision Scores. This can be calculated for various N values and weight values.

$$Geometric\ Average\ Precision\ (N) = exp(\sum_{n=1}^{N} w_n log p_n) \tag{3}$$

Equation 4 is used in the third step to calculate a "Brevity Penalty." With an exponential decline, the shortness penalty penalizes computed translations that are too brief compared to the nearest reference length. The shortness penalty makes up for the absence of a recall phrase in the BLEU score.

$$Brevity\ Penalty = \begin{cases} 1 & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)} & if\ c \leq r \end{cases} \tag{4}$$

Where c is the predicted length and r is the target length.

Then, as illustrated in equation 5, we multiply the Brevity Penalty by the Geometric Average of the Precision Scores to arrive at the Bleu Score.

$$BLEU(N) = Brevity\ Penalty.\ Geometric\ Average\ Precision\ (N) \quad (5)$$

Another metric feature used is the ROUGE algorithm. Recall-Oriented Understudy for Gisting Evaluation is the abbreviation for this method. In essence, it is a collection of measures for assessing machine translations and text summarizations that are done automatically. It contrasts an automatically generated summary or translation with a collection of reference summaries (typically human-produced). The number of "n-grams" that match the text produced by our model and a "reference" is measured by ROUGE-N. A collection of tokens or words is known as an n-gram. A bigram (2 grams) consists of two consecutive words, whereas a unigram (1 gram) only has one word.

6. Materials and methods

Creating a concise, fluid, and, most importantly, excellent description of lengthy text content is known as text summarization. The fundamental goal of information extraction is to be capable of extracting the most important information from a large body of text and displaying it in a human-readable way. Automatic text summarizing techniques could be particularly beneficial as online textual documents increase since more informative material can be viewed quickly. The framework of the knowledge summarization by NLP algorithms is shown in Figure 5. Data were collected from the abstract of the twenty published papers based on High Entropy Alloys available on the Google Scholar database, which has a wide availability of papers compared to other databases. Preprocessing of text has a significant impact on how well models perform. To develop a machine learning model, preprocessing the data is a crucial stage, and the effectiveness of the preprocessing determines the final outcomes. The initial stage of creating a model in NLP is text preparation. In any language used by people, stop words are common. These words are eliminated from our text, allowing us to concentrate more on the crucial information by removing the low-level information they contain. The most significant of these is tokenization. It is the process of segmenting a stream of text information into tokens, which can be words, keywords, sentences, symbols, or other significant components. Numerous open-source programs can be used to tokenize data. Text is divided into sentences through the process of sentence tokenization.

Because the tokenizer is based on a corpus of professional English text, it performs well when applied to literature, reporting, and formal documents. Tokenization is the procedure of dividing text into a sequence of tokens from a string of text. Tokens can be viewed as components, similar to how a word functions as a token in a phrase and how a sentence functions as a token in prose. For evaluating the performance, BLEU and ROUGE metrics features are then implemented. The accuracy of BLEU is slightly increased by using the BLEU-4 feature. Recall, precision, and F1 score are estimated using equations 6, 7, and 8, respectively, and further evaluation of the ROUGE Score.

$$Recall\ Value = \frac{number\ of\ n-grams\ found\ in\ model\ and\ references}{number\ of\ n-grams\ in\ references} \quad (6)$$

$$Precision\ Value = \frac{number\ of\ n-grams\ found\ in\ model\ and\ references}{number\ of\ n-grams\ in\ model} \quad (7)$$

$$F1 - Score = 2 \times \frac{Precision\ Value \times Recall\ Value}{Precision\ Value + Recall\ Value} \quad (8)$$

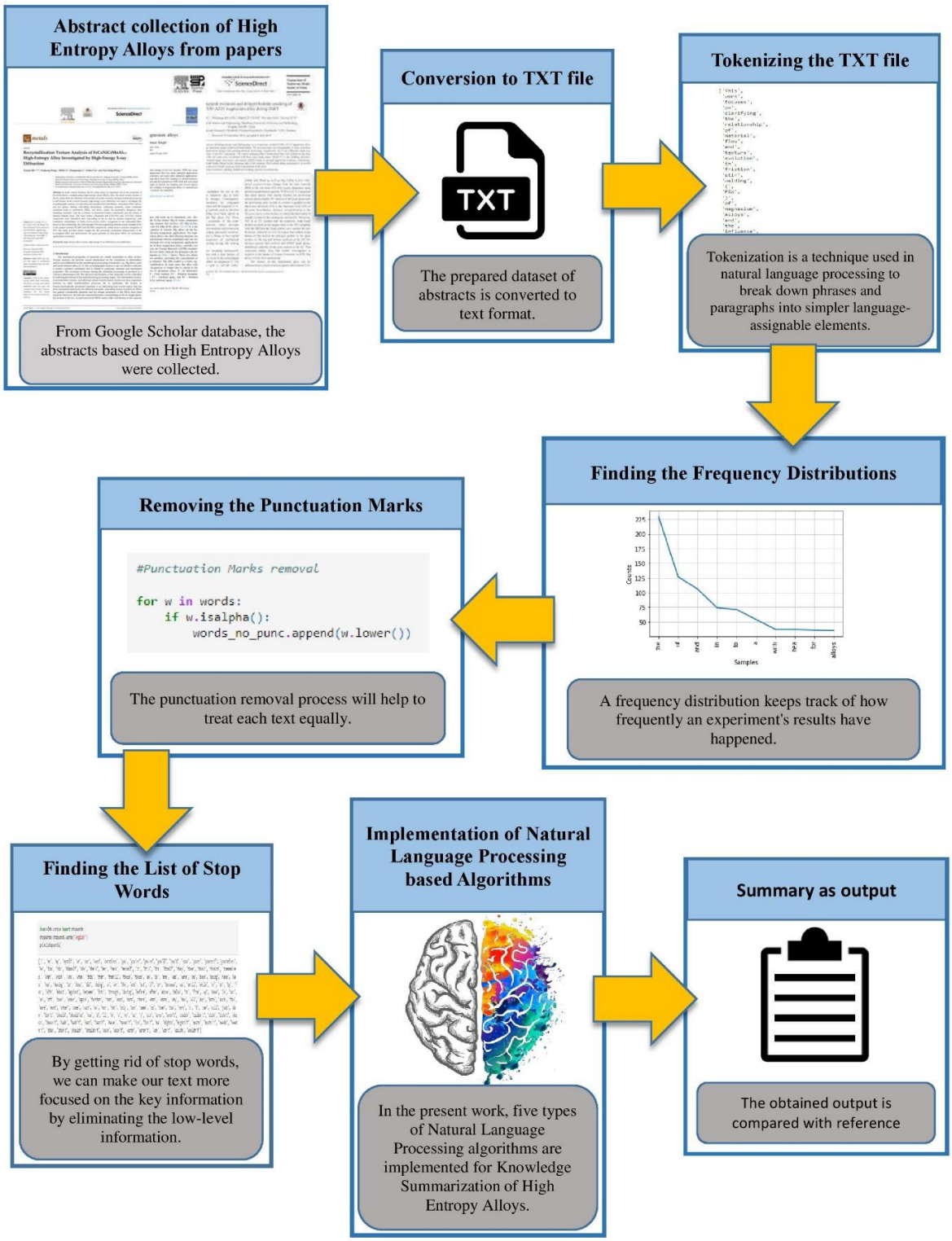
Abstract collection of High Entropy Alloys from papers
From Google Scholar database, the abstracts based on High Entropy Alloys were collected.
Conversion to TXT file
TXT
The prepared dataset of abstracts is converted to text format.
Tokenizing the TXT file
Tokenization is a technique used in natural language processing to break down phrases and paragraphs into simpler language-assignable elements.
Finding the Frequency Distributions
Counts
Samples
A frequency distribution keeps track of how frequently an experiment's results have happened.
Removing the Punctuation Marks
#Punctuation Marks removal
for w in words:
if w.isalpha():
words_no_punc.append(w.lower())
The punctuation removal process will help to treat each text equally.
Finding the List of Stop Words
By getting rid of stop words, we can make our text more focused on the key information by eliminating the low-level information.
Implementation of Natural Language Processing based Algorithms
In the present work, five types of Natural Language Processing algorithms are implemented for Knowledge Summarization of High Entropy Alloys.
Summary as output
The obtained output is compared with reference

Fig. 5. NLP framework used in the recent study.

7. Results

Figure 6 presents the initial plot of the frequency distribution. We can observe that the distribution includes a lot of punctuation and non-content terms like “the,” “of,” “and,” etc. (we refer to them as stop words). Before drawing the graph, we can take them out. To do this, we must import stopwords from the corpus package. A new list of things to ignore is created by adding the stop word list, a list of punctuation, and a list of single figures utilizing + signs. Figure 7 displays the final plot of the frequency distribution after punctuation and stop marks have been removed. Every human language contains a large number of stop words. By eliminating the low-level information, these phrases can be modified to be more concentrated on the important information.

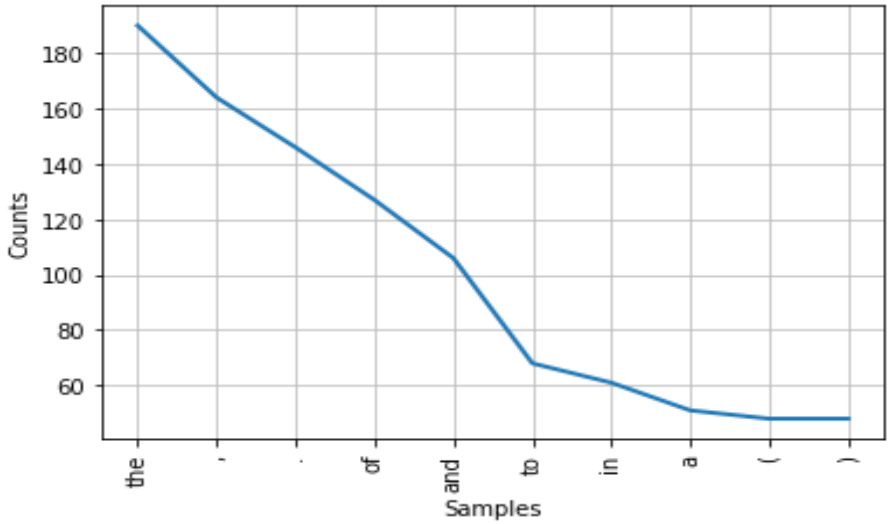

Fig. 6. Initial Frequency Plot.

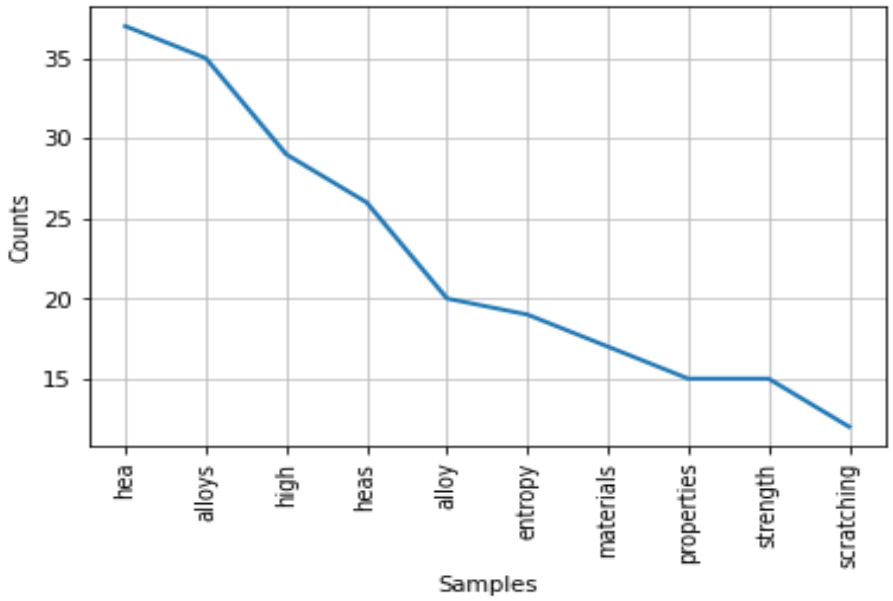

Fig. 7. Final Frequency Plot.

The filtered tokens can alternatively be displayed as a word cloud. This enables us to use the WordCloud().generate from frequencies() method to get a general overview of the corpus. A frequency dictionary containing all the tokens and their frequencies in the text serves as the method's input. It is necessary to first import the Counter package into Python before generating a vocabulary using the input of the filtered text variable. After preprocessing, the word count plot is obtained, as seen in Figure 8. Word in a word cloud is used to visualize text that shows that word's frequency. To highlight important textual information, use a word cloud.

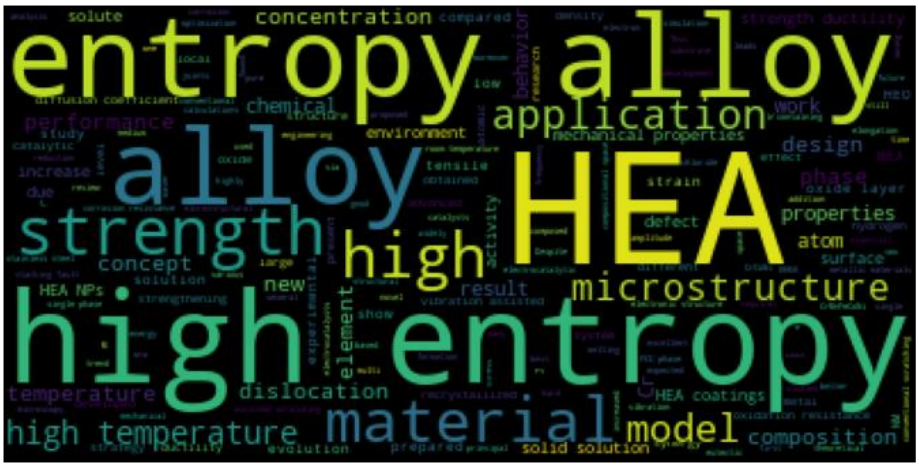

Fig. 8. Word Cloud.

Figure 9 shows the plot of the corresponding algorithms' obtained Individual N- grams values. It is observed that the individual 1-gram value of the LUHN algorithm is the highest in comparison to the other algorithms. Figure 10 shows the obtained BLEU score for each algorithm. To slightly increase the performance of the BLEU score, the BLEU-4 score is obtained, as shown in Figure 11.

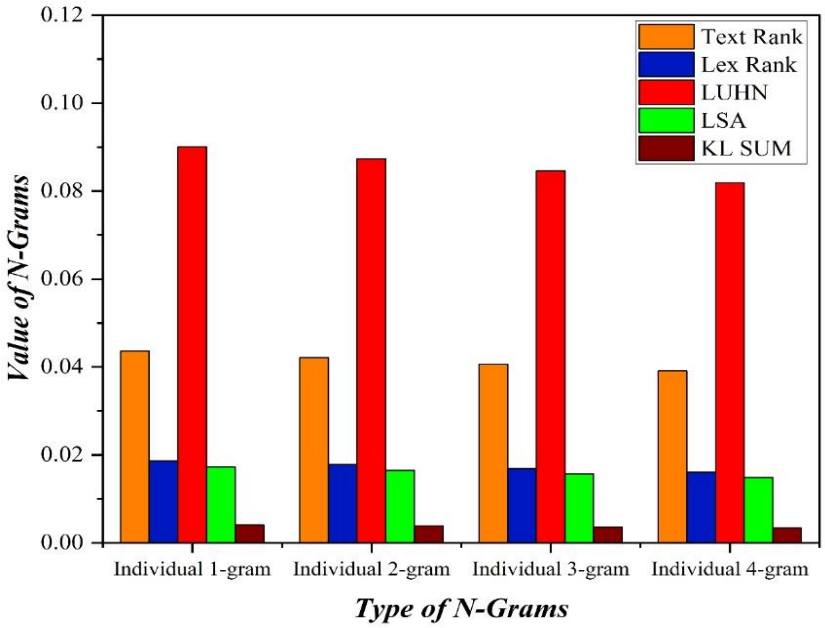

Fig. 9. Obtained N-grams for each NLP algorithms.

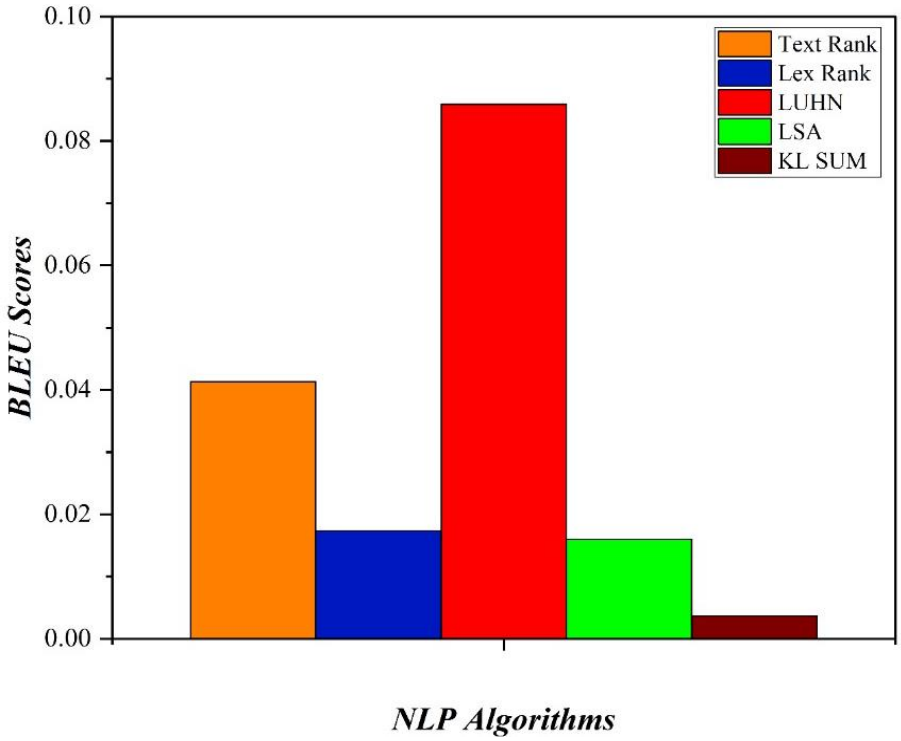

Fig. 10. BLEU Score for the implemented NLP Algorithms.

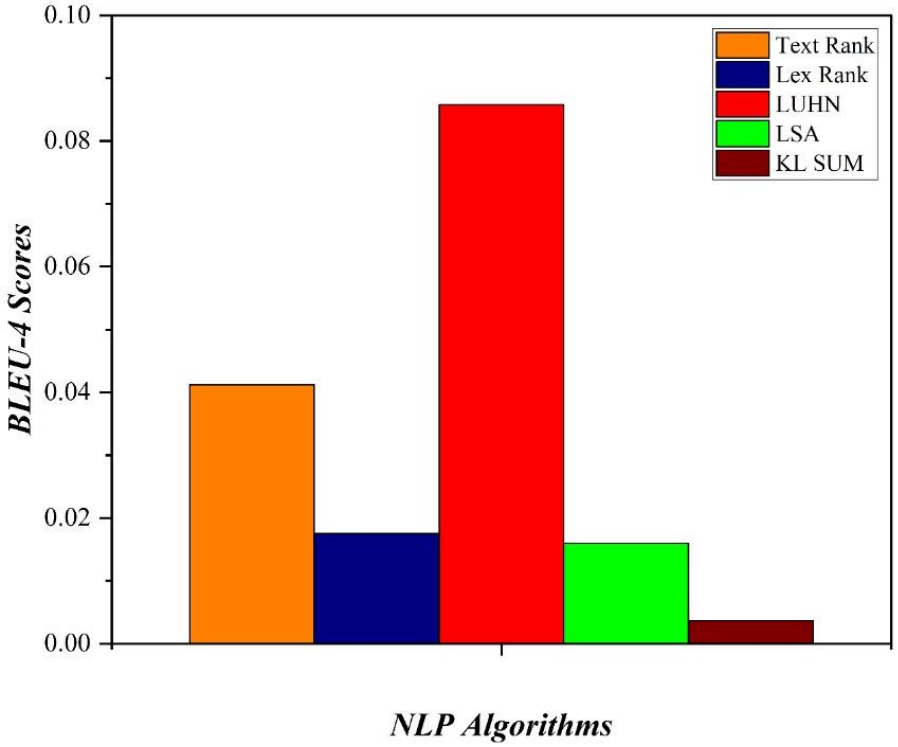

Fig. 11. BLEU-4 Scores for the implemented NLP Algorithms.

Table 1 displays the results of the summative evaluation of the Natural Language Processing algorithms utilized for information summarization using the ROGUE method.

Table 1
NLP implementation results.

NLP Algorithms	Recall	Precision	F1-Score
Text Rank	1.0	0.370	0.540
Lex Rank	1.0	0.299	0.460
LUHN	1.0	0.423	0.595
Latent Semantic Analysis	1.0	0.350	0.519
KL Algorithm	1.0	0.250	0.400

8. Discussion

It is possible to find keywords and the most relevant phrases within a text using a text ranking algorithm based on graphs. It is used to identify the most important sentences in a manuscript by creating a graph with each sentence's vertices being the network's nodes and the edges connecting them being the number of words they share.

With the help of the Pagerank algorithm, the most important sentences are selected based on their phrase structure. We can only use the most important sentences while writing a summary. In the Page Rank algorithm, weights are assigned by Equation 9.

$$X(V_i) = (1-d) + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} X(V_j) \quad (9)$$

Where X(V_i) is the assigned weight of the i webpage, d is the damping factor, In(V_i) is the inbound links of i, and Out(V_j) is the outgoing links of j. To find pertinent terms, the text rank algorithm creates a word network. This network is created by looking at the sequence in which words come before one another. In the text, a connection is made between two words when they are placed next to one another. The connection is more significant if the two terms are placed next to one another more frequently.

The Pagerank method is applied to the resulting network to determine the significance of each word. Despite their relevance, they are all retained except the top third. The relevant terms are arranged into a keywords table as soon as they appear consecutively in the text. We are only permitted to include key sentences when summarizing the content. In order to find pertinent terms, the text rank algorithm creates a word network. This network is constructed by looking at the order in which words come before one another. In the text, a connection is made between two words when they are placed next to one another. The connection is more significant if the two terms are placed next to one another more frequently. The relevance of each word in the resulting network is then determined using the Pagerank algorithm. All but the top third of these words are kept since they are thought to be relevant. If the pertinent words appear right after one another in the text, they are then combined into a keywords table.

The LexRank algorithm assesses the centrality of sentences on graphs as a tool for unsupervised text summarization. The fundamental idea is that sentences "endorse" additional sentences

similar to them for the reader. Because of this, a statement is likely to be significant if it sounds like many others. The unsupervised LexRank technique uses eigenvector centrality-based images to summarize the input text intelligently. A cosine similarity-based adjacency matrix is used to visualize the six required sentences. The sentences are ranked according to how similar they are to the centroid sentence, which serves as the mean for all the assertions in the text file in this method. The N (number of all possible words)-dimensional vectors are represented by the bag of words model, which defines similarity. Equation 10 determines the value of the appropriate dimension in the text's vector representation.

$$Values\ of\ Corresponding\ Dimension = number\ of\ occurrences\ of\ the\ word \times Inverse\ Document\ Frequency\ (IDF) of\ the\ word \tag{10}$$

The associated IDF-modified cosine number of the word is determined by Equation 11.

$$IDF\ Modified\ Cosine\ Value\ (x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2} \sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}} \tag{11}$$

where idf_w is the inverse document frequency of a word w and tf_w,s is the number of times the word w appears in the sentences.

By analyzing sample corpora of natural literature, latent semantic analysis (LSA) is a computational method for computer simulation and modeling of the definition of words and passages. LSA can accurately approximate many facets of learning and interpreting human languages. It supports a range of applications in information retrieval, instructional technology, and other information processing challenges where complicated wholes can be viewed as additive functionalities of constituent elements. Latent Semantic Analysis, commonly known as LSI (for Latent Semantic Indexing), simulates the contribution of words combined into coherent sequences to natural language. It uses the well-known Singular Value Decomposition (SVD) method from matrix algebra, which wasn't until the late 1980s, and the development of potent digital computing systems and algorithms to make use of them that it was possible to apply to such complicated events. To create a semantic space for a speech, LSA divides a sizable representational text corpus into rectangular cells, each containing a transform of the frequency of a given the word in a specific section.

The Luhn algorithm uses a TF-IDF-based methodology. According to their frequency, it only chooses words with tremendous significance. The words appearing at the document's beginning are given more weight. The initial step is identifying which words contribute most to the document's meaning. A frequency analysis must first be conducted, then placing significant but not insignificant English terms.

In the second step, the terms that appear most often in the document are determined, and a subset of those that are less common but still important is then chosen. Calculating a sentence's grade requires the usage of Equation 12.

$$Score = \frac{(Number\ of\ meaningful\ words)^2}{Span\ of\ meaningful\ words} \tag{12}$$

The Luhn approach makes creating summaries out of a group of words simple. The technique can be applied in two steps. First, we attempt to determine which keywords are most crucial to understanding the text as a whole. In order to accomplish this, according to Luhn, a frequency study should be done first, then phrases that contain major but not inconsequential English terms must be found. A subset of the less frequent but still significant terms is then selected after the most frequent terms in the material are identified in the second step.

Every time KL divergence decreases, a sentence is added to the summary using the Kull-back-Lieber (KL) method, which would be predicated on the greedy method strategy. The KL approach decides whether or not it detracts from its own intake vocabulary by reducing the summary vocabulary. If there are L elements in a synthesized and D for a particular text, the KL algorithm offers a criterion for selecting summaries. Although various error metrics are in use, sending as little information as possible is our top priority. These two models reduce our problem to a finite number of variables. The best method for deciding which distribution is better is to use the Kullback-Liber algorithm to find which distribution maintains the most information from the original data source.

The findings imply that phrase centrality ratings derived from graphs are used by the LexRank algorithm for unsupervised text summarization. The fundamental tenet is that statements "recommend" other assertions to readers if they are comparable to their own. A remark is likely extremely important if it is virtually identical to multiple others. The significance of the word "recommending" in this statement helps to illustrate its significance. As a result, many other phrases must be linked to a statement for it to communicate and be included in the summary. This makes perfect sense and allows applying the technique to any brand-new content.

According to reports, the LexRank algorithm's f1-score and accuracy value is lower than the LSA approach's. Latent Semantic Analysis (LSA) is used to understand textual information statistically. The foundation of topic modeling is this tactic. A specific topic and data matrix of our current arrangements and materials will be split into two separate matrices to achieve the main purpose.

The Luhn Algorithm has higher f1-scores than the LexRank and LSA algorithms; it has been found. The Luhn technique's Term Frequency-Inverse Document Frequency (TF-IDF) basis may also be shown. Only the most important terms are selected based on how consistently they are used. A phrase's first few words are given additional weight. The Text rank method likewise follows the f1-score of the Luhn algorithm. It is a fairly evident objective for Text Rank to identify how closely related each expression is to each other phrase in the text. It is observed that in addition to the ROUGE score, the BLEU-4 score of the LUHN algorithm is higher than the other algorithms.

9. Conclusion

The assistance of machine learning and natural language processing (NLP) has paved the way for generating autonomous text summaries. Natural language processing holds immense significance in facilitating effective interactions between machines and humans. Ongoing research efforts

continue to propel the development of NLP, as its potential is far from fully realized. Within this context, integrating NLP and automatic knowledge summarization has emerged as a vital approach for condensing extensive and cumbersome papers across diverse domains, including technology, finance, law, medicine, and literature. Both business and academia stand to reap substantial benefits from these techniques, with promising prospects for future advancements.

Drawing on the latest research, the following significant conclusions can be derived:

- High Entropy Alloys research publications underwent evaluation using five distinct NLP techniques: Text Rank, KL Algorithm, LSA, Luhn, and Lex Rank.
- The findings indicate that the Luhn algorithm exhibited the highest f1-score of 0.595 compared to the other algorithms, showcasing its superior performance in knowledge summarization tasks.
- The BLEU-4 scores substantiate the results obtained from the ROUGE metric features, with the Luhn algorithm consistently demonstrating a higher score than the other algorithms, affirming its effectiveness in producing quality summaries.
- It is crucial to acknowledge the limitation concerning accuracy scores, as the size of the datasets utilized can influence them. Expanding the dataset size presents an opportunity for improving accuracy and mitigating this limitation.
- Future research endeavors should focus on implementing more advanced algorithms such as BERT, XLNet, and GPT2 for knowledge summarization purposes, enabling the exploration of more sophisticated summarization capabilities.

Funding

This research received no external funding.

Conflicts of interest

The authors declare no conflict of interest.

Authors contribution statement

AM, VSJ, VM: Conceptualization; DD, EMS,AD: Data curation; AM,VM,AD: Formal analysis; DD,VM.VKJ,AM: Investigation; AM,DD,VM,VSJ,SD,EMS: Methodology; VSJ: Project administration; AM,VSJ,VM: Resources; AM,VM,EMS: Software; VSJ: Supervision; VM,AM: Validation; AM, AD, EMS: Visualization; AM, EMS: Roles/Writing – original draft; AM,EMS,VSJ: Writing – review & editing.

References

[1] Chowdhary KR. Fundamentals of artificial intelligence. Springer; 2020.

[2] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Informatics Assoc 2011;18:544–51. https://doi.org/10.1136/amiajnl-2011-000464.

[3] Chopra A, Prashar A, Sain C. Natural language processing. Int J Technol Enhanc Emerg Eng Res 2013;1:131–4.

[4] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."; 2009.

[5] Lehnert WG, Ringle MH. Strategies for natural language processing. Psychology Press; 2014.

[6] Olivetti EA, Cole JM, Kim E, Kononova O, Ceder G, Han TY-J, et al. Data-driven materials research enabled by natural language processing and information extraction. Appl Phys Rev 2020;7. https://doi.org/10.1063/5.0021106.

[7] Xie L, Zhu X, Sun W, Jiang C, Wang P, Yang S, et al. Investigations on the material flow and the influence of the resulting texture on the tensile properties of dissimilar friction stir welded ZK60/Mg–Al–Sn–Zn joints. J Mater Res Technol 2022;17:1716–30. https://doi.org/10.1016/j.jmrt.2022.01.127.

[8] Mohammadi E. Knowledge mapping of the Iranian nanoscience and technology: a text mining approach. Scientometrics 2012;92:593–608. https://doi.org/10.1007/s11192-012-0644-6.

[9] Maghrebi M, Waller ST, Sammut C. Text mining approach for reviewing the ready mixed concrete literature. 2nd Int. Conf. Civ. Build. Eng. Informatics, Univ. Osaka, Tokyo, Japan, 2015, p. 105–9.

[10] Liu G, Boyd M, Yu M, Halim SZ, Quddus N. Identifying causality and contributory factors of pipeline incidents by employing natural language processing and text mining techniques. Process Saf Environ Prot 2021;152:37–46. https://doi.org/10.1016/j.psep.2021.05.036.

[11] Sathiyamoorthi P, Kim HS. High-entropy alloys with heterogeneous microstructure: Processing and mechanical properties. Prog Mater Sci 2022;123:100709. https://doi.org/10.1016/j.pmatsci.2020.100709.

[12] Zhang Y, Wang D, Wang S. High-Entropy Alloys for Electrocatalysis: Design, Characterization, and Applications. Small 2022;18. https://doi.org/10.1002/smll.202104339.

[13] Dewangan SK, Mangish A, Kumar S, Sharma A, Ahn B, Kumar V. A review on High-Temperature Applicability: A milestone for high entropy alloys. Eng Sci Technol an Int J 2022;35:101211. https://doi.org/10.1016/j.jestch.2022.101211.

[14] Zhang Q, Zhang S, Luo Y, Liu Q, Luo J, Chu PK, et al. Preparation of high entropy alloys and application to catalytical water electrolysis. APL Mater 2022;10. https://doi.org/10.1063/5.0097479.

[15] Xue L, Ding Y, Pradeep KG, Case R, Castaneda H, Paredes M. Development of a non-equimolar AlCrCuFeNi high-entropy alloy and its corrosive response to marine environment under different temperatures and chloride concentrations. J Alloys Compd 2022;928:167112. https://doi.org/10.1016/j.jallcom.2022.167112.

[16] Wu Q, He F, Li J, Kim HS, Wang Z, Wang J. Phase-selective recrystallization makes eutectic high-entropy alloys ultra-ductile. Nat Commun 2022;13:4697. https://doi.org/10.1038/s41467-022-32444-4.

[17] Oliveira JP, Shen J, Zeng Z, Park JM, Choi YT, Schell N, et al. Dissimilar laser welding of a CoCrFeMnNi high entropy alloy to 316 stainless steel. Scr Mater 2022;206:114219. https://doi.org/10.1016/j.scriptamat.2021.114219.

[18] Wu D, Kusada K, Nanba Y, Koyama M, Yamamoto T, Toriyama T, et al. Noble-Metal High-Entropy-Alloy Nanoparticles: Atomic-Level Insight into the Electronic Structure. J Am Chem Soc 2022;144:3365–9. https://doi.org/10.1021/jacs.1c13616.

[19] Rong Z, Wang C, Wang Y, Dong M, You Y, Wang J, et al. Microstructure and properties of FeCoNiCrX (X Mn, Al) high-entropy alloy coatings. J Alloys Compd 2022;921:166061. https://doi.org/10.1016/j.jallcom.2022.166061.

[20] Wei D, Gong W, Wang L, Tang B, Kawasaki T, Harjo S, et al. Strengthening of high-entropy alloys via modulation of cryo-pre-straining-induced defects. J Mater Sci Technol 2022;129:251–60. https://doi.org/10.1016/j.jmst.2022.04.055.

[21] Yang Y, Li Z, Zhang W, Ma Y, Xiong Q, Li W, et al. Concentration of "Mysterious Solute" in CoCrFeNi high entropy alloy. Scr Mater 2022;211:114504. https://doi.org/10.1016/j.scriptamat.2022.114504.

[22] Jin Z, Zhou X, Hu Y, Tang X, Hu K, Reddy KM, et al. A fourteen-component high-entropy alloy@oxide bifunctional electrocatalyst with a record-low Δ E of 0.61 V for highly reversible Zn–air batteries. Chem Sci 2022;13:12056–64. https://doi.org/10.1039/D2SC04461G.

[23] Banko L, Krysiak OA, Pedersen JK, Xiao B, Savan A, Löffler T, et al. Unravelling Composition–Activity–Stability Trends in High Entropy Alloy Electrocatalysts by Using a Data-Guided Combinatorial Synthesis Strategy and Computational Modeling. Adv Energy Mater 2022;12. https://doi.org/10.1002/aenm.202103312.

[24] Wetzel A, von der Au M, Dietrich PM, Radnik J, Ozcan O, Witt J. The comparison of the corrosion behavior of the CrCoNi medium entropy alloy and CrMnFeCoNi high entropy alloy. Appl Surf Sci 2022;601:154171. https://doi.org/10.1016/j.apsusc.2022.154171.

[25] Zhou X-Y, Zhu J-H, Wu Y, Yang X-S, Lookman T, Wu H-H. Machine learning assisted design of FeCoNiCrMn high-entropy alloys with ultra-low hydrogen diffusion coefficients. Acta Mater 2022;224:117535. https://doi.org/10.1016/j.actamat.2021.117535.

[26] Ostovari Moghaddam A, Sudarikov M, Shaburova N, Zherebtsov D, Zhivulin V, Ashurali Solizoda I, et al. High temperature oxidation resistance of W-containing high entropy alloys. J Alloys Compd 2022;897:162733. https://doi.org/10.1016/j.jallcom.2021.162733.

[27] Wei D, Wang L, Zhang Y, Gong W, Tsuru T, Lobzenko I, et al. Metalloid substitution elevates simultaneously the strength and ductility of face-centered-cubic high-entropy alloys. Acta Mater 2022;225:117571. https://doi.org/10.1016/j.actamat.2021.117571.

[28] Yuan JL, Wang Z, Jin X, Han PD, Qiao JW. Ultra-high strength assisted by nano-precipitates in a heterostructural high-entropy alloy. J Alloys Compd 2022;921:166106. https://doi.org/10.1016/j.jallcom.2022.166106.

[29] Hou JX, Liu SF, Cao BX, Luan JH, Zhao YL, Chen Z, et al. Designing nanoparticles-strengthened high-entropy alloys with simultaneously enhanced strength-ductility synergy at both room and elevated temperatures. Acta Mater 2022;238:118216. https://doi.org/10.1016/j.actamat.2022.118216.

[30] Doan D-Q, Fang T-H. Effect of vibration parameters on the material removal characteristics of high-entropy alloy in scratching. Int J Mech Sci 2022;232:107597. https://doi.org/10.1016/j.ijmecsci.2022.107597.

[31] Dhungana DS, Mallet N, Fazzini P-F, Larrieu G, Cristiano F, Plissard SR. Self-catalyzed InAs nanowires grown on Si: the key role of kinetics on their morphology. Nanotechnology 2022;33:485601. https://doi.org/10.1088/1361-6528/ac8bdb.

[32] Sadek D, Dhungana DS, Coratger R, Durand C, Proietti A, Gravelier Q, et al. Integration of the Rhombohedral BiSb(0001) Topological Insulator on a Cubic GaAs(001) Substrate. ACS Appl Mater Interfaces 2021;13:36492–8. https://doi.org/10.1021/acsami.1c08477.

[33] Dhungana DS, Grazianetti C, Martella C, Achilli S, Fratesi G, Molle A. Two-Dimensional Silicene–Stanene Heterostructures by Epitaxy. Adv Funct Mater 2021;31. https://doi.org/10.1002/adfm.202102797.

[34] Dhungana DS, Hemeryck A, Sartori N, Fazzini P-F, Cristiano F, Plissard SR. Insight of surface treatments for CMOS compatibility of InAs nanowires. Nano Res 2019;12:581–6. https://doi.org/10.1007/s12274-018-2257-8.

[35] Sefene EM, Tsegaw AA, Mishra A. Process Parameter Optimization of 6061AA Friction Stir Welded Joints Using Supervised Machine Learning Regression-Based Algorithms. J Soft Comput Civ Eng 2022;6:127–37. https://doi.org/10.22115/scce.2022.299913.1350.

[36] Bonaventura E, Dhungana DS, Martella C, Grazianetti C, Macis S, Lupi S, et al. Optical and thermal responses of silicene in Xene heterostructures. Nanoscale Horizons 2022;7:924–30. https://doi.org/10.1039/D2NH00219A.

[37] Zhang Y. History of High-Entropy Materials. High-Entropy Mater., Singapore: Springer Singapore; 2019, p. 1–33. https://doi.org/10.1007/978-981-13-8526-1_1.

[38] Yeh J -W., Chen S -K., Lin S -J., Gan J -Y., Chin T -S., Shun T -T., et al. Nanostructured High-Entropy Alloys with Multiple Principal Elements: Novel Alloy Design Concepts and Outcomes. Adv Eng Mater 2004;6:299–303. https://doi.org/10.1002/adem.200300567.

[39] Miracle DB. High entropy alloys as a bold step forward in alloy development. Nat Commun 2019;10:1805. https://doi.org/10.1038/s41467-019-09700-1.

[40] Ye YF, Wang Q, Lu J, Liu CT, Yang Y. High-entropy alloy: challenges and prospects. Mater Today 2016;19:349–62. https://doi.org/10.1016/j.mattod.2015.11.026.

[41] Dada M, Popoola P, Adeosun S, Mathe N. High Entropy Alloys for Aerospace Applications. Aerodynamics, IntechOpen; 2021. https://doi.org/10.5772/intechopen.84982.

[42] Cui L, Ma B, Feng SQ, Wang XL. Microstructure and Mechanical Properties of High-Entropy Alloys CoCrFeNiAl by Welding. Adv Mater Res 2014;936:1635–40. https://doi.org/10.4028/www.scientific.net/AMR.936.1635.

[43] Lippold JC, Kiser SD, DuPont JN. Welding metallurgy and weldability of nickel-base alloys. John Wiley & Sons; 2011.

[44] Vendan SA, Gao L, Garg A, Kavitha P, Dhivyasri G, Rahul SG. Interdisciplinary treatment to arc welding power sources. Springer; 2019.

[45] Kong X, Yang Q, Li B, Rothwell G, English R, Ren XJ. Numerical study of strengths of spot-welded joints of steel. Mater Des 2008;29:1554–61. https://doi.org/10.1016/j.matdes.2007.12.001.

[46] Chen S, Tong Y, Liaw PK. Additive Manufacturing of High-Entropy Alloys: A Review. Entropy 2018;20:937. https://doi.org/10.3390/e20120937.

[47] Xuan Li R, Zhang Y. Entropic Alloys for Cryogenic Applications. Stainl. Steels Alloy., IntechOpen; 2019. https://doi.org/10.5772/intechopen.82351.

[48] Geanta V, Voiculescu I. Characterization and Testing of High-Entropy Alloys from AlCrFeCoNi System for Military Applications. Eng. Steels High Entropy-Alloys, IntechOpen; 2020. https://doi.org/10.5772/intechopen.88622.